AD-A061 012 DAYTON UNIV OHIO RESEARCH INST F/G 17/2

PRELIMINARY EVALUATION OF THE CXC SYSTEM AS AN AUTOMATIC SPEECH--ETC(U)

OCT 78 D G LEET F33615-77-C-0510

UNCLASSIFIED UDR-TR-78-79 AMRL-TR-78-85 NL

1 OF 1

AD A061012

END

DATE FILMED

1-79

DDC

LEVEL II

AMRL-TR-78-85

1

AD-A061012

PRELIMINARY EVALUATION OF THE CxC SYSTEM AS AN AUTOMATIC SPEECH RECOGNIZER

Duane G. Leet

OCTOBER 1978

Approved for public release; distribution unlimited

AEROSPACE MEDICAL RESEARCH LABORATORY
AEROSPACE MEDICAL DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433

78 10 30 033

NOTICES

TECHNICAL REVIEW AND APPROVAL

AMRL-TR-78-85

This report has been reviewed by the Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Henning E. von Gierke

HENNING E. VON GIERKE
Director
Biodynamics and Bioengineering Division
Aerospace Medical Research Laboratory

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AMRL-TR-78-85	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* PRELIMINARY EVALUATION OF THE CxC SYSTEM AS AN AUTOMATIC SPEECH RECOGNIZER		5. TYPE OF REPORT & PERIOD COVERED Final Report 1 March 1977 - 1 May 1978
		6. PERFORMING ORG. REPORT NUMBER UDR-TR-78-79
7. AUTHOR(s) Duane G. Leet		8. CONTRACT OR GRANT NUMBER(s) F33615-77-C-0510
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Dayton Research Institute 300 College Park Avenue Dayton, Ohio 45469		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2312-V4-15 61102F
11. CONTROLLING OFFICE NAME AND ADDRESS Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio 45433		12. REPORT DATE October 1978
		13. NUMBER OF PAGES 41
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

ACCESSION for	
NTIS	White Section X
DDC	Buff Section ☐
UNANNOUNCED	☐
JUSTIFICATION	
BY	
DISTRIBUTION/AVAILABILITY CODES	
Dist. AVAIL. and/or SPECIAL	
A	

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Speech Recognition System
Auditory System Model
Signal Processing
Neuromine Networks
Analog Cochlea

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The CxC System is a general purpose hybrid hardware and software system for extracting transient sequential patterns from continuous analog, quasi-periodic, audiofrequency signals in real- or near real-time. The design of the system relies heavily on studies of the information processing characteristics of the mammalian auditory system.

This report presents an overview of the CxC System configured for automatic speech recognition and a preliminary performance evaluation of the system in this capacity.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

PREFACE

This is the final report on work performed under Contract F33615-77-C-0510, Work Unit 2312V415, "Auditory Analysis of Specific Acoustic Techniques," University of Dayton Research Institute (UDRI), Dayton, Ohio; Dr. Duane G. Leet, Principal Investigator.

The entire program was conducted in support of Project No. 2312, Task 2312V4, "Applications of Basic Biological Principles and Mechanisms to Operation and Design of Air Force Systems: administered by the Biological Acoustics Branch of the Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio. The period covered is 1 March 1977 to 1 May 1978. Dr. Thomas J. Moore was the initiator and monitor of this research.

Air Force support of this program is justified by the following arguments. First, the human auditory system is a speech processing system that far surpasses current technological capabilities. Study of this system from a rigorous information-theoretic viewpoint could lead to improvement of existing technology or development of an entirely new technology that can be used by the Air Force in intelligence, reconnaissance, air traffic control, air mission control, and man-machine interface. Second, the voiced speech signal is comprised of a sequence of glottal pulses. Mathematically, a glottal pulse can be considered to be a short sample of a Fourier transformable function embedded in noise. This noise comes from two sources. The statistics of one source can vary with time but are independent of the glottal pulse. The statistics of the other source depends on the sequence of glottal pulses preceding the given pulse; that is, the glottal pulse frequency characteristics are context-sensitive.

The human auditory system is able, therefore, to recognize sequences of context-sensitive signals embedded in noise, and to do this in real time. Again, the capabilities of this system far surpass current technological capabilities to process such signals. The development of electronic systems with these capabilities would have application both to the general signal recognition problem in Air Force intelligence and to electronic warfare.

In addition to this report, three other reports were published with the support of this contract:

Leet, Duane G. (1977), "An Efficient Computer Algorithm to Determine All Partitions of a Sequence of Length n Into Subsequences of Length m, m $\leq$ n," UDRI-TR-77-32, University of Dayton, Research Institute, Dayton, Ohio.

Cashin, John L., Jr. (1978), "A Review of Techniques Utilized in Studying Varying Neural Aggregates of the Guinea Pig Auditory System," UDR-TR-78-25, University of Dayton Research Institute, Dayton, Ohio.

Leet, Duane, G. and Christopher W. Walsh (1978), "The CxC System: Standard Operating Procedures for the Analysis of Synthetic and Real Speech, Revision II," UDR-TR-77-58, University of Dayton Research Institute, Dayton, Ohio.

The author expresses his sincere appreciation to Dr. Moore and to John L. Cashin, Jr., Captain Vincent Mortimer, Dr. J. Ryland Mundie, and Christopher Walsh for their invaluable assistance.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

SECTION 1
INTRODUCTION

Research in speech recognition and associated technology is of interest to a number of United States Department of Defense Agencies, including the following Air Force organizations (Beek, et al, 1977):

- Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio
- Air Force Avionics Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio
- Rome Air Development Center, Griffiss Air Force Base, Rome, New York
- Air Force Office of Scientific Research, Arlington, Virginia
- Air Force Electronic Systems Division, Bedford, Massachusetts

The Aerospace Medical Research Laboratory (AMRL), in particular, has had a research program related to speech recognition technology for over 16 years. As initially conceived, the objective of the research program was not, however, specifically to develop a speech recognition system; it was to study the auditory system from a psychoacoustic, biomedical, and neurophysiological point of view and to apply the principles learned to Air Force-oriented problems. The primary medium in which knowledge of the auditory system stored is the CxC (C-square) System, which is an evolving hardware/ software model of the peripheral auditory system.

Table 1 is a list, modified from Beek, et al (1977), of Air Force problems that may benefit from speech recognition technology. Early in the development of CxC, before the implementation of Fast Fourier Transform, Linear Predictive Coefficients, and Large Scale Integration, it was proposed as the front end to a speech recognition system. A limited evaluation of CxC in this capacity has been made. This report summarizes these results.*

---

*Captain Donald B. Warmuth has also designed a speech recognition system using CxC. His results are described elsewhere (Warmuth, 1978).

TABLE 1

AIR FORCE PROBLEMS THAT MAY BENEFIT FROM SPEECH RECOGNITION TECHNOLOGY[1]

Secruity Applications

- Verification or rejection of an individual based on his speech patterns (This situation is characterized by controlled context, a controlled environment, a controlled communication system, and a cooperative speaker.)
- Identification of an individual based on his speech patterns (This application lacks the above simplifying conditions.)
- Determination of the emotional state of the speaker (e.g. stress effects)
- Recognition of spoken codes
- Surveillance of communication channels
  - Recognition of a keyword or a set of keywords embedded in narrow band conversational speech as expected from a radio link
  - Language identification

Command and Control Applications

- System control (aircraft flight control, fire control, navigation, electronic warfare, tactical situation displays, crew training, etc.)
- Voice-operated computer input/output (each telephone a terminal)
- Data handling and record control
- Material handling (logistics)
- Remote control (dangerous material)
- Administrative record control

Data Transmission and Communication

- Vocoder systems (the real-time transformation of a speech analog waveform into a parametric representation)
- Speech synthesis (the generation of a speech analog waveform from parametric data such as those generated by a vocoder system)
- Bandwidth reduction or, more generally, bit-rate reduction of a speech transmission
- Ciphering, coding, and scrambling of a speech transmission

Processing Distorted Speech

- Diver (Helium) speech-like transformations (generally deterministic mapping of speech characteristics)
- Astronaut communication (degraded speech.)
- Stressed speech (relatively short-term distortion of speech characteristics)

Processing Speech-Like Signals

- Radar return classification (e.g. classification of vehicle type)
- Radar signal classification in an electronic warfare situation

---

[1]Adapted from Beek, et al (1977)

Section 2 provides a model of the speech recognition problem from the point of view of communication theory. Section 3 describes the CxC system without regard to application. Section 4 describes the organization of the automatic speech recognition front end that includes CxC. Section 5 describes the preliminary evaluation that has been performed on this latter system.

SECTION 2

A MODEL OF THE SPEECH RECOGNITION PROBLEM

Modeled as the standard communication channel of Figure 1, the speech recognition problem can be described in terms of three components, a transmitter, which generates a message and transmits a signal, a channel, through which the signal passes and is usually degraded, and a receiver, which extracts its best estimate of the message from the signal using technology that today is comprised of some combination of decision-theoretic and syntactic pattern recognition techniques.

The conceptual design of the transmitter is dictated in part by the receiver's conceptual design. Referring to the block diagram in Figure 2, the message- a sentence- is generated by the Pattern Sequence Generator. It is comprised of a sequence of patterns (words) chosen from a Pattern Dictionary. The Pattern Sequence Generator is assumed to have an appropriate amount of "intelligence" to create a meaningful message consistent with constraints contained in a set of Pattern Sequence Rules. The message is passed to the Pattern Element Generator, which uses a deterministic Pattern-to-Element Map to transform the message into a pattern element sequence (a sequence of phonemes). This form of the message is, in turn, modified by the Pattern Element Sequence Modifier. This component uses a set of deterministic Element Sequence Modification Rules that delete or substitute elements, particularly at the boundary between two patterns (word boundaries). Finally, the message is passed to the Signal Generator. The Signal Generator model's structure is shown in Figure 3. The sequence of pattern elements from the Element Sequence Modifier is processed by a Controller that generates coefficients for a time-varying filter, amplitude values, and controls a switch that can provide either an impulse train or noise as input to the filter.

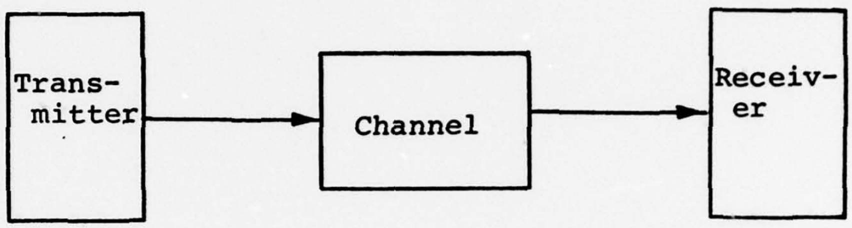

Figure 1. The Standard Communication Channel.

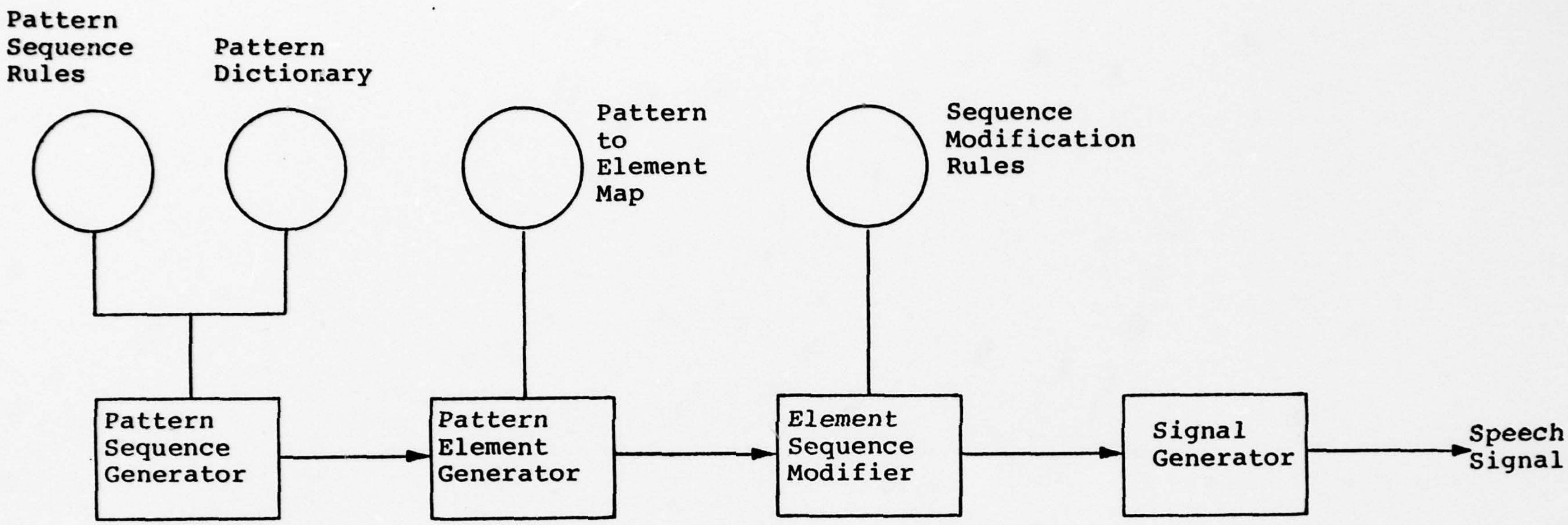

Figure 2. The Transmitter Model Block Diagram.

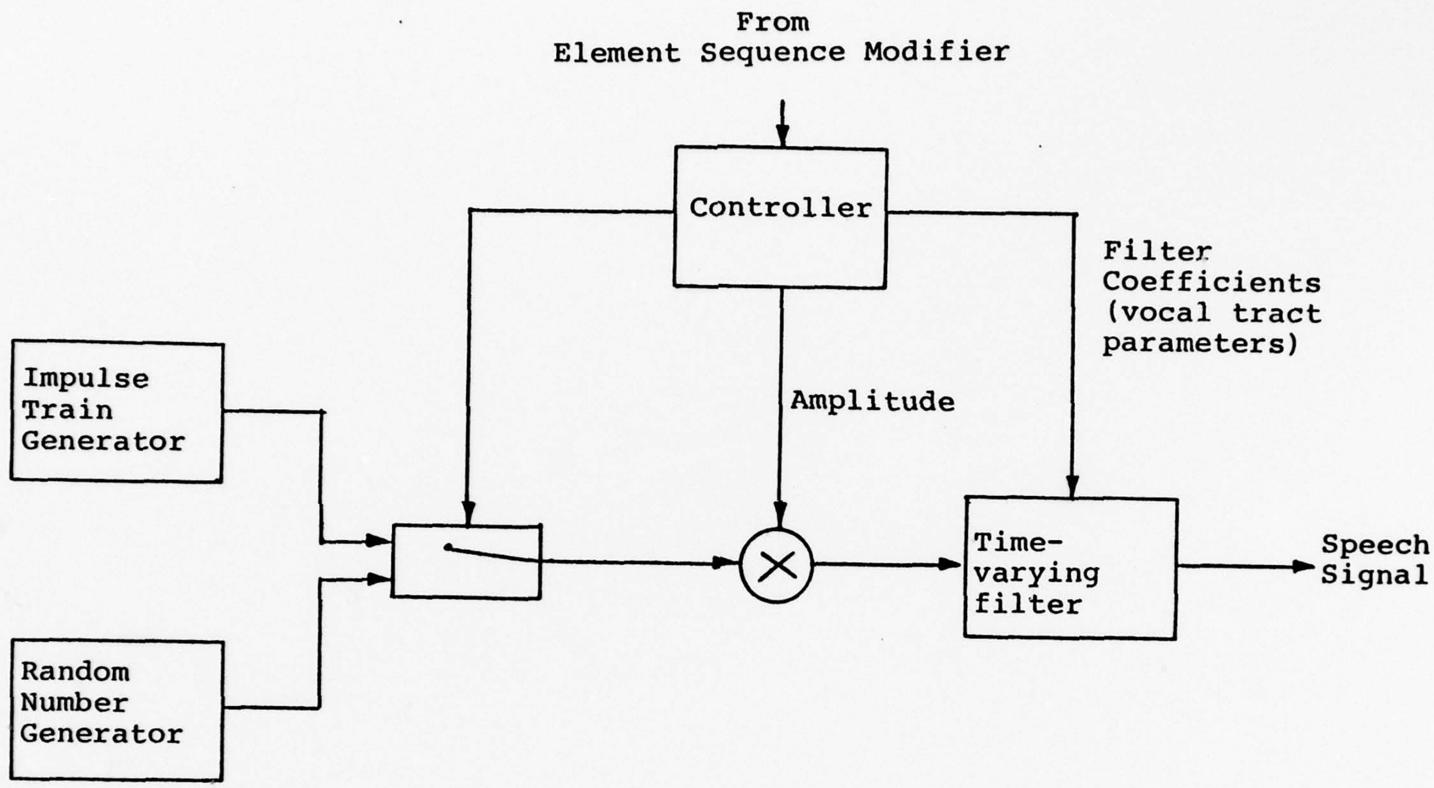

Figure 3. The Signal Generator Model.

The Receiver design, as proposed by Klatt (1977) and slightly modified for this discussion, is summarized in Figure 4. The "speech input" is a transformed version of the signal generated by the Transmitter, modified by both noise and linear filters. It is processed by the Front End component, which partitions the signal into segments and generates a multi-dimensional parameter vector for each segment. A priori information about the sentences that can be spoken and how they might be spoken is contained in the Acoustic Segment Lexical Decoding (ACLD) Network. In a separate off-line one-shot generation, the speaker was asked to speak about twenty sentences carefully chosen to contain all the different segment acoustic forms that could occur during the speaker's conversation. Parameter vectors are formed from these segments and placed in a special reference file, the Diphone Dictionary. Each node in the ACLD Network points to a parameter vector in the reference file.

The Search Strategy component within the Bottom End section parses the acoustic segment sequence using the ACLD Network. The a priori probabilities dictate the order in which node exit paths are tested. Each destination node reference pattern is compared to the next acoustic pattern vector using a distance measure. A decision function combines the a priori and a posteriori information to select the best path through the Network. When an end-of-sentence mark is detected in the Network, the candidate sentence is sent to the Search Strategy component in the Top End section. It uses an Augmented Transition Network with Semantics to decide whether the sentence is in acceptable form. If it is, the sentence is output; otherwise, the Bottom End Search Strategy component is notified. It locates the next best path through the Network, starting from the beginning of the acoustic segment sequence, and this is evaluated by the Top End section. This process is repeated until an acceptable sentence is identified or the Network possibilities are exhausted.

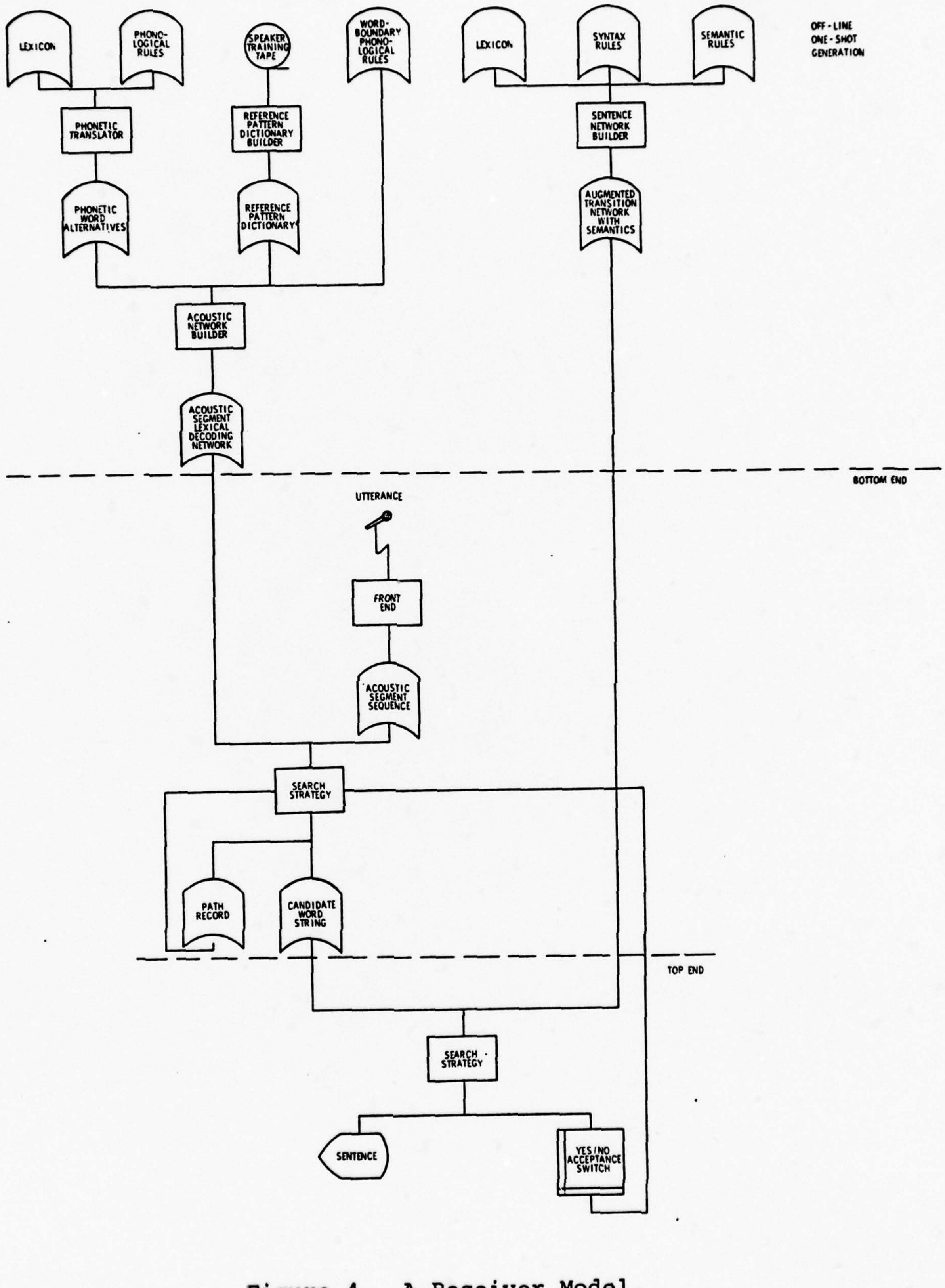

Figure 4. A Receiver Model.

SECTION 3
OVERVIEW OF CxC DESIGN

The CxC System is a general purpose hybrid hardware and software system for extracting transient sequential patterns from continuous analog, audio-frequency signals in real- or near real-time. The design of the system relies heavily on studies of the information processing characteristics of the mammalian peripheral auditory system. A hardware block diagram of CxC is shown in Figure 5. The analog input signal first passes through an electronic analog device called the Cochlear Filter. As a model, the Cochlear Filter's input signal is analogous to the sound pressure at the input to the middle ear. Its Middle Ear circuit is an active bandpass filter centered at about 2 kHz and with a bandwidth of about 3.5 kHz. The Analog Cochlea is a transmission line comprised of a cascade of up to 48 (depending on the model used) low pass filters. Each filter is tapped to provide an output voltage analogous to the displacement at a point on the cochlea's basilar membrane (See Stewart (1972) for a comprehensive technical description.). The effective magnitude Bode function from the Cochlear Filter's input to a channel output is that of an asymmetrical bandpass filter, with the low frequency skirt having a 6 dB/octave slope and the high frequency skirt having a 100 dB/octave slope. The channel electronically closest to the input, channel 1, corresponds to the basal end of the cochlea and has the highest peak frequency; the channel electronically furthest from the input, channel 48, corresponds to the apical end of the cochlea and has the lowest peak frequency. Each channel has been tuned to provide a maximum gain of one from input to channel output. That is, any sinusoid input with a frequency in the audio range should appear on one of the output terminals unattenuated. As with the cochlea, a signal introduced at the Analog Cochlea's input requires a significant amount of time, varying logarithmically with distance, to travel the length of the transmission line, with the circuit components selected to approximate the times found in the cochlea

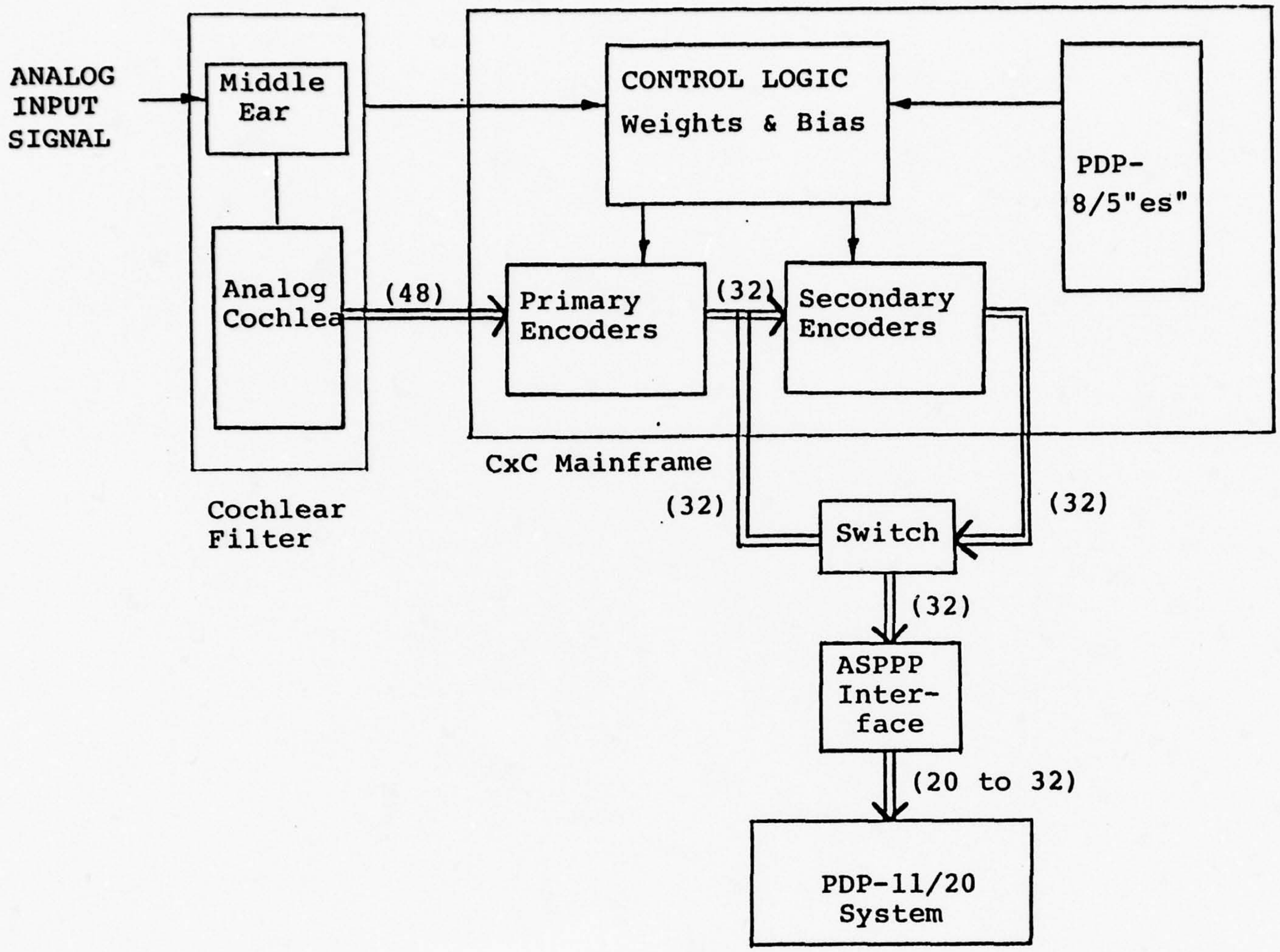

Figure 5. The CxC System Hardware Block Diagram.

(See Mundie, et al (1974) for details.). An important consequence of this delay is the ability of the Analog Cochlea to store about 1.75 cycles of signal within it, regardless of signal frequency.

The overall response of the Cochlear Filter to a speech signal is illustrated in Figure 6. Note that the higher frequency components of the signal are extracted by the low-numbered channels, while the lower frequency components are extracted by the high-numbered channels. Also note the signal delay as a function of channel.

The CxC Mainframe is a unique piece of hardware that uses a neuromime network to extract and digitize features from the Cochlear Filter output. The Primary and Secondary Encoder sections are each comprised of a rank of syncoders, supported by synapse button, sample and hold, and feedback circuits. A detailed hardware description of these sections can be found in Mundie et al (1978). The conceptual aspects of the design will be emphasized in this report.

The syncoder neuromime is a summing function and leaky integrator followed by a comparator. The comparator's reference is a time-dependent, exponentially decaying "threshold function". Whenever the comparator's input equals the threshold function, the syncoder produces a standard digital logic pulse, the threshold function is reset to a high value, and the syncoder becomes unresponsive for a period of time (the refractory period), after which the threshold function begins its exponential decay again. The syncoder's transfer function is, therefore, input signal dependent.

The synapse button is an electronic component designed to interface syncoders.* It has two inputs and a single output that can be connected to (fanout) up to eight syncoders. The synapse button is basically a switch that is closed whenever the pulse output of the syncoder is high. The input voltage supplied to the

*The description of the synapse button, sample and hold, switch, and ASPPP circuits are taken in part from Warmuth (1978).

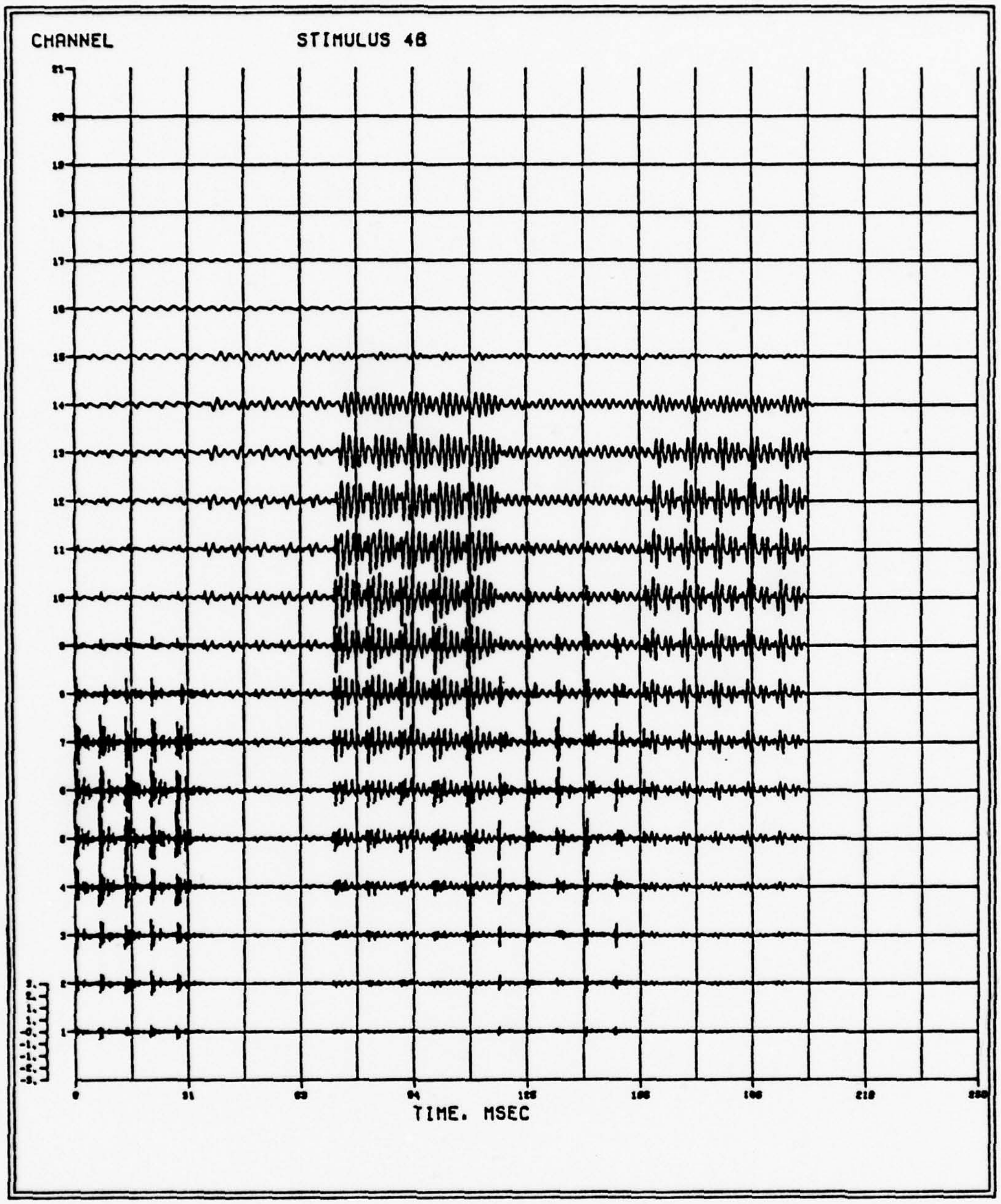

Figure 6. The Cochlear Surface Generated by the Cochlear Filter in Response to a Sequence of Five Vowels Extracted from Real Speech.

switch is provided by the sample and hold circuits. The voltage produced at the synapse button output exponentially increases toward the input voltage whenever the switch is closed and immediately begins to exponentially decay whenever the switch is opened.

The sample and hold (s&h) circuits supply the voltage sources required by the synapse buttons and DC levels that are added at the syncoder summing junctions to bias the various time-varying signals. These circuits are controlled by the PDP-8/S digital computer. The PDP-8/S addresses each s&h board individually and supplies a predetermined voltage to that s&h board through a digital to analog converter. The PDP-8/S requires less than two minutes to sequentially address all 1728 s&h boards in CxC. The voltage on a s&h board is about 97% of its original value at the end of this time.

There are 32 syncoders in the primary encoder level. Two of these are in special circuits that will be described in the next two paragraphs. Each of the remaining syncoders is embedded in a circuit connected to a different Cochlear Filter output and adjusted (integration time, threshold decay constant, refractory time, biasing, and feedback) so that it will fire on each peak of the highest frequency that can reach that output at maximum input amplitude. Descriptions of the behavior of these syncoders to various inputs can be found in Mundie (1969), Ziskin and Mundie (1971), Rock (1973), and Warmuth (1978).

One syncoder in the primary encoder level functions as a pitch period marker: a pulse on this channel indicates that a voiced sound is present and its position in time marks the approximate beginning of the pitch period. This special syncoder circuit accomplishes this by first low-pass filtering the signal to 300 Hz. A voltage proportional to the average amplitude is generated by full wave rectification and integration of the filtered signal and this voltage is used to bias the syncoder. The filtered signal is also presented to a second input of the syncoder. This circuit causes the syncoder to fire on the large amplitude peaks that occur at the beginning of each impulse excitation.

A second syncoder in the primary encoder level has a pulse rate that is logorithmically proportional to the amplitude of the signal input to the Cochlear Filter. This is accomplished by presenting a short interval average of the signal to the syncoder. A characteristic of the syncoder's behavior is that it responds at a rate logorithmically proportional to the level of a DC input.

The second encoder level also has 32 outputs, but it has only 30 syncoders; the amplitude and pitch period lines from the first level are passed through untouched and unused. The syncoders serve as coincidence detectors between the pulse patterns of two first level syncoders. The two first level syncoders are always either one-half (five cochlea channels) or one wavelength (eight cochlea channels) apart and serve as peak-to-valley or peak-to-peak detectors respectively. The range of wavelengths a syncoder is sensitive to is adjustable, mainly through the synapse buttons.

To complete the description of Figure 5, the Switch component is a 2 position, 32 pole electronically controlled switch that switches either the Primary or Secondary Encoder output to the ASPPP interface. ASPPP is an acronym for Asynchronous Pulse Pattern Processor. It is a hardware interface used to sample the 32 digital outputs from the Switch and pass the results to the PDP-11/20 digital computer for storage and later processing. Each five microseconds the ASPPP looks for up to two rising edges of pulses on the 32 channels, starting with the first channel. If it finds at least one, it records the channel(s) and the time since the last pulse was recorded on any channel. Although the ASPPP can only record the first two pulses it encounters in a five microsecond sweep, no significant loss of data appears to result.*

*In an informal inspection of speech data by Warmuth (1978), it was found that two channels had fired simultaneously less than 5% of the time.

SECTION 4

PROPOSED ORGANIZATION OF CxC AS THE FRONT END TO A SPEECH RECOGNITION SYSTEM

The classic implementation of a decision-theoretic pattern recognition system consists of a Preprocessor to condition the raw signals, a Feature Extractor to extract the relevant information and generally reduce the dimensionality of the problem, and a Classifier, to provide either a definitive classification, an probabilities or the probabilities themselves. Following this structure, Figure 7 is the proposed block diagram for CxC as a speech recognition system.

A Preamp circuit, the Middle Ear circuit, and the Analog Cochlea within the Cochlear Filter comprise the Preprocessor. Both syncoder levels, and the Staging section of a Correlator component comprise the Feature Extractor. The Correlator section of the Correlator, a Scoring program, and two stages of discrimination programs comprise the Classifier.

Four new system components have been introduced. Conceptually, the Staging section is a bank of up to twenty 1000 bit shift registers that hold the most recent 10 msec of CxC response data. The Correlator has the capacity to store up to 20 reference patterns. A reference pattern is a "snapshot" of the Staging section contents obtained during a training session. A reference pattern can be considered to be either a 20 (channel) x 1000 (10 microsecond time unit) binary array or a 20,000-dimensional binary vector. The correlation computation performed is almost the inner product of each reference pattern with whatever is in the Staging section. The word "almost" is included because of the existence of a "smear" option. When this option is operating, the effect is the same as replacing an equal number of "0"s on either side of every one in the reference pattern with "1"s before the standard inner product operation is performed.

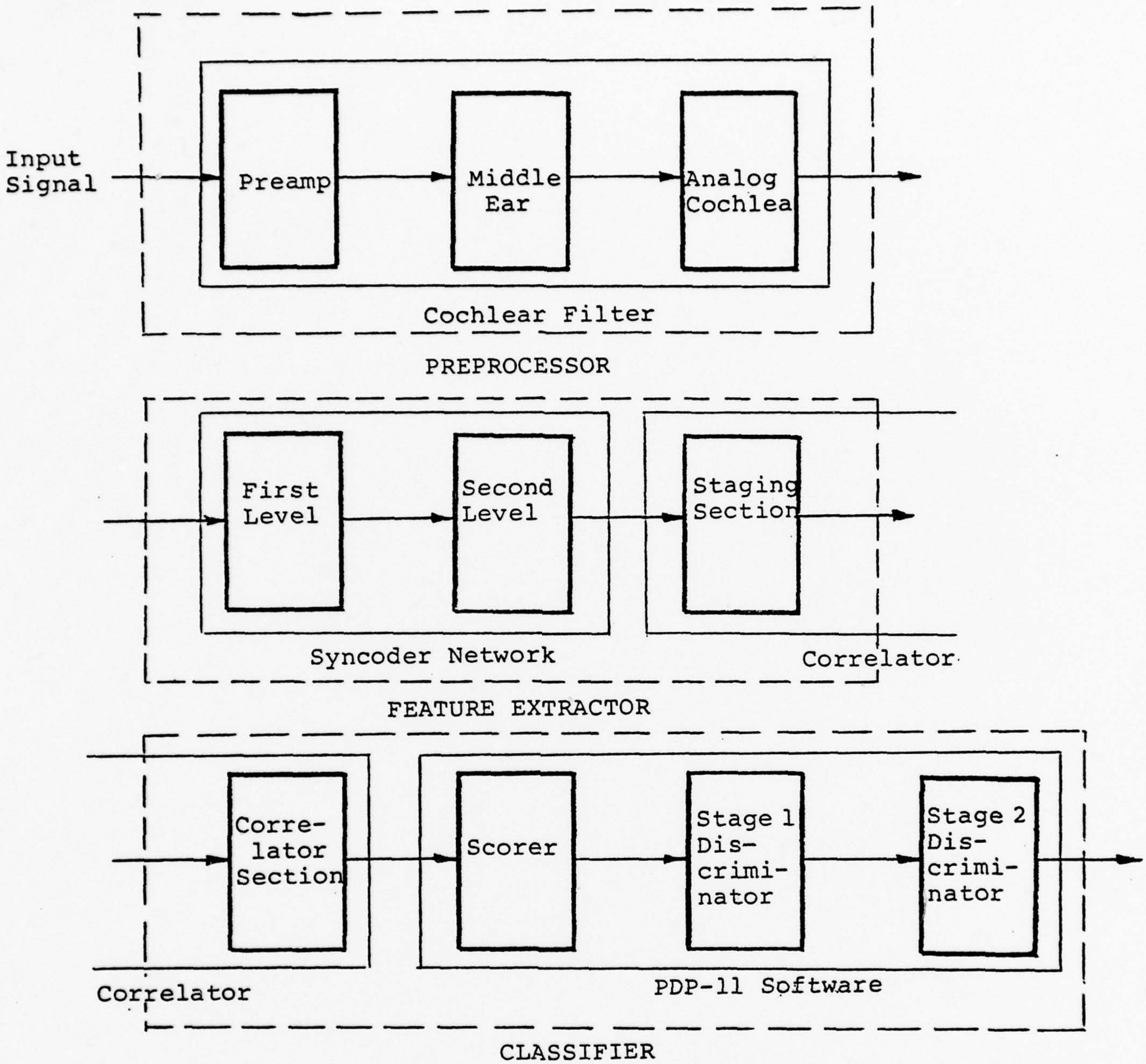

Figure 7. The CxC System Block Diagram for Speech Recognition Applications. The Functional Components are Defined by the Heavy Lines, the Hardware Components by the Light Lines, and the Basic Information-Theoretic Components by the Dashed Lines.

A principal assumption supporting the analysis of the CxC response is that the CxC transformation provides a unique, detectable trajectory for each phoneme. This research effort is oriented toward finding the optimum design for this discrimination. (Hardware constraints limit the number of CxC output channels that can be processed simultaneously. Thus, to be completely accurate, this assumption must be modified by including the word "selected" before "CxC transformation".)

Another assumption is that the unique phoneme trajectories should be stored as reference patterns and that the CxC response should be compared to them. It was realized early in the development that the direct implementation of this philosophy-ie. storing the entire trajectories as reference patterns and performing some kind of running correlation with the CxC output pattern- was technically not feasible. Therefore, a two-step approach was taken. The reference patterns stored in the system are about a pitch period in length. Furthermore, this is done for only a subset of the phonemes, the 20 or less mentioned above. Time dependent (10 processed by the Stage 1 Discriminator to provide measures of similarity for each reference pattern for each pitch period. In summary, the (CxC output) x (time) binary space, where CxC output is 20-dimensional and time is in 5 microsecond units, has been transformed into a (reference pattern similarity scores) x (pitch period) real or integer space, where (reference pattern similarity scores is 20-dimensional or less and (pitch period) is a subset of integers. The Stage 2 Discriminator contains phoneme trajectory reference patterns in this latter space, which are compared to the incoming patterns. The assumption, of course, is that with all the information reduction that has occurred, discrimination can still occur.

SECTION 5

DEVELOPMENT AND EVALUATION OF THE SCORER AND THE STAGE 1 DISCRIMINATOR

5.1 CHRONOLOGY FOR THE DEVELOPMENT AND EVALUATION OF CxC AS A SPEECH RECOGNITION SYSTEM FRONT END

The proposed chronology for the development and evaluation of CxC as a speech recognition system front end has these identified major tasks:

- Task 1: Develop a speech synthesis-by-rule system
- Task 2: Develop a correlator
- Task 3: Develop a software system to display the output of CxC, to extract and store reference patterns in a PDP-11 disk file, and to manipulate and display the reference patterns
- Task 4: Develop and evaluate the Scorer and the Stage 1 Discriminator
- Task 5: Develop and evaluate the Stage 2 Discriminator

The first three tasks have been completed and the systems developed are described in other reports [Warmuth (1976), White (1977), Hartrum (1978), and Leet and Walsh (1977)].

Table 2 is the expanded chronology for Task 4. This section summarizes the work performed through Task 4.4. Data are available on several 4.5 subtasks, but at the termination of the contract they had not been thoroughly analyzed.

5.2 DEVELOPMENT OF THE SCORER DESIGN

The design philosophy for the CxC Classifier requires the comparison of the contents of the Correlator's Staging section with reference patterns that are, in turn, obtained from the Staging section. The measure used for comparison and the rationale behind its development are presented in this section.

TABLE 2

SUBTASKS FOR TASK 4: DEVELOP AND EVALUATE THE SCORER AND THE STAGE 1 DISCRIMINATOR

4.1 Develop and Implement the Scorer Design

- 4.1.1 Theoretical Development
- 4.1.2 Program Development (DTLST Program)

4.2 Preliminary Evaluation of Scorer Design

- 4.2.1 Selection of Reference Patterns
- 4.2.2 Generation of a Vowel-Vowel Continuous Synthetic Speech Stimulus (The Vowel-Vowel Transition Stimulus QQ)
- 4.2.3 Acquisition of Reference Patterns
- 4.2.4 Evaluation of Scorer Design Using the Vowel-Vowel Transition Stimulus

4.3 Develop and Implement the Stage 1 Discriminator Design

- 4.3.1 Theoretical Development
- 4.3.2 Program Development (CHANNEL and PICKSP programs)

4.4 Preliminary Evaluation of Stage 1 Discriminator Design

- 4.4.1 Test Design on Vowel-Vowel Transition Stimulus
- 4.4.2 Design Modifications Where Necessary

4.5 Evaluation and Characterization of System Performance (Through State 1 Discriminator) Against the Remaining Synthetic Phonemes

- 4.5.1 Nasals and Semivowels
- 4.5.2 Diphthongs
- 4.5.3 Fricatives (Voiced and Unvoiced)
- 4.5.4 Stops (Voiced and Unvoiced)
- 4.5.5 Aspirants and Affricates
- 4.5.6 Balanced Word Lists

Under the perfect conditions of (1) a deterministic synthesizer, (2) a speech input of sustained, isolated, synthetic phonemes from the set of reference phonemes, (3) unvarying recording conditions, and (4) no noise, perfect classification of input phonemes should be possible using a Hamming distance measure, which counts the number of mismatches in two binary patterns. The decision function is trivial: the input phoneme is the one whose Hamming distance is zero. Of course, this assumes the feature extractor has provided separated reference patterns.

If any restriction is lifted that permits the Staging section pattern to not exactly match any reference pattern, then a comparative measure is required that has the metric property that the lower the score the closer one pattern is to another. The Hamming distance is such a candidate. It can be easily computed from the Correlator's outputs. If $N(\underline{r}_i)$ is the number of ones in a reference pattern $\underline{r}_i$, $N(\underline{c}(n\Delta t))$ is the number of ones in the Staging section, and $M_i(n\Delta t)$ is the number of matches generated by the Correlator section, then the Hamming distance is

$$H_i(n\Delta t) = N(\underline{r}_i) + N(\underline{c}(n\Delta t)) - 2\, M_i(n\Delta t).$$

There is, however, an aspect of the Hamming distance's interpretation of "closeness" that is unappealing. Suppose the Staging section contents are $\vec{c} = (1\ 1\ 1\ 1\ 0\ 0\ 0)$ and the reference patterns are $\vec{r}_1 = (1\ 1\ 0\ 0\ 0\ 0\ 0)$ and $\vec{r}_2 = (1\ 1\ 1\ 1\ 1\ 1\ 1)$. Then the Hamming distances are $H(\vec{r}_1) = 2$ and $H(\vec{r}_2) = 3$. Therefore, the phoneme associated with $\vec{r}_1$ would be selected as the most likely input phoneme. But, from another point of view, 50% (2 mismatches/4 dimensions) of the dimensions containing ones are different for $\vec{r}_1$, while 43% (3 mismatches/7 dimensions) are different for $\vec{r}_2$. Somehow this latter measure of closeness seems more appropriate. Formalizing this concept, in terms of the data available from the Correlator Section, the Mismatches/Dimensions measure, is given by

$$S_i(n\Delta t) = \frac{H_i(n\Delta t)}{N(\underline{r}_i) + N[\underline{c}(n\Delta t)] - M_i(n\Delta t)}$$

This measure has not been shown to satisfy the triangle inequality condition for a metric. However, no counterexamples could be generated to prove it was not a metric.

The Scorer section, implemented as the DTLST program on the PDP-11/20 by White (1977), computes the Mismatches/Dimensions Measure for all reference patterns in each 10 microsecond time interval. Figure 8 is a typical page of print generated by DTLST. Note that the measure is expressed as an integer with a value between 0 and 100. The user can input a threshold value, 85 in this sample; a line of print is generated only if a score at or below threshold occurred during a time interval, and only those scores at or below threshold are printed.

5.3 THE STAGE 1 DISCRIMINATOR DESIGN

The Stage 1 Discriminator is provided with a Mismatches/Dimensions score for each reference pattern for each ten microsecond time increment as the input signal is shifted through the correlator. This component reduces the data to one score per reference pattern per time segment, where a time segment is either an input signal pitch period, or, if no pitch period marker is present, a standard period, such a ten milliseconds. The score generated for each reference pattern is the lowest score occurring during the time segment. The Discriminator can generate any one of three kinds of output: (1) it can provide the scores for each reference pattern, (2) it can provide the reference pattern with the lowest score, or (3) it can provide the reference pattern with the lowest score and that score.

The Stage 1 Discriminator is implemented on the PDP-11/20 as two programs. CHANNEL reads the original CxC output file and generates a file listing the times of occurrence of the pitch period marker, which is found in the first channel of the CxC output. PICKSP uses this file and the output of the Correlator to generate time segment (pitch period) data.

PATTERN FAMILY NAME IS SYNTHL INPUT FILE TYPE IS PEAK SCORES

FUNCTION MISMATCHES/DIMENSIONS
THRESHOLD 85

IDS	OW	UU	IY	II	OO	EE	AE	ER	AA	UH	NN	MM	NG	LL	RR	##	D1	D2	D3
POINTS	12	13	28	28	14	22	20	18	19	7	35	19	24	10	14				
TIME(MS)																			
9.79	.	.	.	.	.	.	59	.	.	.	.	.	.	.	.	18			
10.83	.	.	.	.	.	.	82	.	.	.	.	.	.	.	.	20			
11.24	.	.	.	.	.	84	.	.	.	.	.	.	.	.	.	22			
11.86	.	.	.	.	.	.	.	.	.	.	.	.	84	.	.	19			
12.79	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	18			
15.78	.	.	.	.	.	83	.	.	.	.	.	.	.	.	.	19			
16.34	.	.	.	.	.	82	.	.	.	.	.	.	.	.	.	18			
16.50	.	.	.	82	.	.	.	.	.	.	.	.	.	.	.	18			
17.06	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	21			
17.82	.	.	85	.	.	.	.	.	.	.	.	.	.	.	.	18			
19.12	.	.	83	.	.	.	.	.	.	.	.	.	.	.	.	14			
23.20	.	.	.	.	.	81	.	.	.	.	.	.	.	.	.	16			
23.21	.	.	.	.	.	.	80	.	.	.	.	.	.	.	.	16			
24.86	.	.	.	.	.	.	37	.	.	.	.	.	.	.	.	19			
25.26	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	18			
30.63	.	.	.	.	.	.	80	.	.	.	.	.	.	.	.	16			
32.28	.	.	.	.	.	.	33	.	.	.	.	.	.	.	.	20			
32.68	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	18			
38.06	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	17			
39.37	.	.	.	.	.	.	36	.	.	.	.	.	.	.	.	21			
40.11	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	18			
45.49	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	18			
46.61	.	.	.	.	.	84	.	.	.	.	.	.	.	.	.	22			
46.79	.	.	.	.	.	.	38	.	.	.	.	.	.	.	.	22			
47.54	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	19			
52.91	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	18			
54.04	.	.	.	.	.	84	.	.	.	.	.	.	.	.	.	22			
54.21	.	.	.	.	.	.	38	.	.	.	.	.	.	.	.	22			
54.96	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	19			
60.33	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	17			
61.64	.	.	.	.	.	.	36	.	.	.	.	.	.	.	.	21			
62.38	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	18			
67.74	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	17			
69.01	.	.	.	.	.	.	36	.	.	.	.	.	.	.	.	21			
69.79	.	.	62	.	.	.	.	.	.	.	.	.	.	.	.	18			
75.83	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	17			
76.47	.	.	.	.	.	.	36	.	.	.	.	.	.	.	.	21			
77.58	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	18			
82.57	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	17			
83.88	.	.	.	.	.	.	36	.	.	.	.	.	.	.	.	21			
84.62	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	18			
90.66	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	18			
91.29	.	.	.	.	.	.	38	.	.	.	.	.	.	.	.	22			
92.04	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	19			
98.08	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	17			
98.71	.	.	.	.	.	.	36	.	.	.	.	.	.	.	.	21			
99.83	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	18			
105.85	.	.	.	.	.	.	80	.	.	.	.	.	.	.	.	16			
106.08	.	.	.	.	.	.	40	.	.	.	.	.	.	.	.	20			
107.23	.	.	82	.	.	.	.	.	.	.	.	.	.	.	.	18			
112.90	.	.	.	.	.	.	81	.	.	.	.	.	.	.	.	18			
113.33	.	.	.	.	.	84	.	.	.	.	.	.	.	.	.	22			

1

Figure 8. Typical Page of DTLST Output for MISMATCHES/ DIMENSIONS Function.

An example page of output from PICKSP is shown in Figure 9. The starting time of the pitch period is listed in the first column, "Start PP". The second column, "Pat", contains a list of the reference pattern IDs. The next three columns are under the label "Group Analysis". A group is a set of pitch periods in sequence. The number of pitch periods in a group is defined by the user at the beginning of the run. From the listing's viewpoint, if the number of pitch periods in a group is 6, then the group consists of the present pitch period and the previous five pitch periods. The entries in the three columns under the Group Analysis heading, "Min", "Avg", and "Max" provide the minimum of all the minimum scores in the group, the average of the minimum scores in the group, and the maximum of the minimum scores in the group respectively.

The "Min" column next to the "Group Analysis" columns contains the minimum score that occurred during the pitch period for each reference pattern. The entries under the heading "Times of Occurrence (From Beginning of PP)" define the times that the minimum occurred relative to the beginning of the pitch period. Up to 10 entries are possible.

5.4 PRELIMINARY EVALUATION OF THE STAGE 1 DISCRIMINATOR DESIGN

5.4.1 The QQ Test Stimulus

A preliminary evaluation of the CxC automatic speech recognizer design through the Stage 1 Discriminator has been made using a specially constructed synthetic vowel sequence called QQ, which was generated by the AUTOTA synthesis-by-rule program. In the notation of Appendix A, the vowel sequence was OO, IY, OO, AA, OO, OW, OO, AE, OO, AE, OW, AE, AA, AE, ER, AE, IY, AE, IY, AA, IY, OO, IY, IR, IY, OW, IY, OW, OO, OW, AE, OW, ER, OW, AA, IY, AA, OO, AA, AE, AA, OW, AA, UH, AA, ER, AE, ER, AA, ER, OW, ER, OO, ER, IY, and ER. AUTOTA provides an exponential transition from one vowel

START PP	PAT	GROUP ANALYSIS MIN	AVG	MAX	MIN	PP ANALYSIS TIMES OF OCCURRENCE (FROM BEGINNING OF PP)									
181.92	OW	88	91.6	98	92	0.01	0.02	0.03	0.04	0.05					
	UU	93	95.0	98	93	3.32	3.33								
	IY	67	72.6	81	69	0.02									
	II	85	91.0	94	94	0.99									
	OO	91	93.0	96	91	1.30	1.31	1.32							
	EE	93	94.0	95	94	3.39	5.06	5.07	5.08	5.09	5.10				
	AE	92	92.8	94	94	6.02									
	ER	92	93.6	94	94	1.33	1.34	1.35	1.36	3.34	3.35	3.36	3.37	3.38	
	AA	94	97.0	88	98	0.94	0.95	0.96	0.97	0.99	1.30	1.31	1.32	1.33	1.34
	UH	85	94.6	97	97	1.35	1.36	1.37	3.38	3.39	3.40	5.05	5.06	5.07	5.08
	NN	94	95.2	96	96	0.96	0.97	0.99							
	MM	91	93.4	98	91	5.05	5.06	5.07							
	NG	95	96.2	98	97	0.94									
	LL	92	95.8	98	92	5.10									
	RR	96	97.6	98	98	4.72	4.73	4.74	4.75						
188.35	OW	86	89.2	92	86	0.07									
	UU	93	94.2	96	94	3.32									
	IY	62	68.8	73	62	6.41									
	II	85	89.2	94	85	0.07	0.08	0.09	0.10						
	OO	91	92.4	93	93	1.29	1.30	1.31	4.74						
	EE	94	94.4	95	95	3.37	3.38	5.06	5.07	5.08	5.09				
	AE	92	93.2	94	94	5.05	5.06	5.07	5.08	6.01	6.02				
	ER	92	93.6	94	94	1.32	3.32	3.34	3.35	3.36	3.37				
	AA	94	97.2	98	98	1.29	1.30	1.31	1.32	1.33	1.34	6.03	6.04	6.05	6.06
	UH	85	94.6	97	97	0.09	0.10	0.32	1.34	1.35	4.75	5.05	5.06	5.07	5.08
	NN	94	95.6	97	97	6.40	6.41								
	MM	91	92.6	94	94	5.05									
	NG	95	96.4	98	96	0.07	0.08	0.09	0.10	0.32					
	LL	92	94.8	98	92	5.08									
	RR	96	97.6	98	98	4.71	4.72	4.73	4.74	4.75					

Figure 9. Sample Page of PICKSP Output.

to the next. As demonstrated in the upper part of Figure 9, since the length of the utterance chosen for each vowel was relatively short, the length of time the formants of the vowels were at their target values (Appendix B) was relatively short; i.e., most of the vowel sample was of a vowel in transition from the previous vowel.

QQ was presented to CxC using the REFPAT program, under the SAMPLE option. REFPAT created a file on the PDP-11 disk containing the response of CxC to QQ.

5.4.2 Reference Patterns

Dr. J. Ryland Mundie selected the phonemes used for the reference patterns based on his extensive experience with CxC and speech characteristics. The phonemes chosen were the vowels, the nasals, and the semivowels /l/ and /r/. The REFPAT program, under the REFERENCE PATTERN option, was used to extract the CxC responses to sustained synthetic versions of the phonemes. The reference patterns were extracted such that the response to an entire pitch period fell within a 6.95 msec window, which was the size selected for the Staging section. Figure 10 is an example reference pattern.

5.4.3 Correlation

The correlation was performed by the software correlator CORR, written by Pat White (1977). This program simulates the operation of the hardware correlator, but at a speed hundreds of times slower. In fact, about eighteen hours were required to correlate about half of the response to QQ, or about 15,000 times real time.

5.4.4 Stage I Performance

The DTLST program was run on the files generated by CORR, and the PICKSP program, guided by the CHANNEL program output file, was run on the DTLST output files. A typical example of Stage I's performance is shown in the lower half of Figure 11. Summarizing the results, for all vowels in QQ, a significant portion of the pitch periods during the time when a vowel was the target was scored (lowest score) as being generated by the target

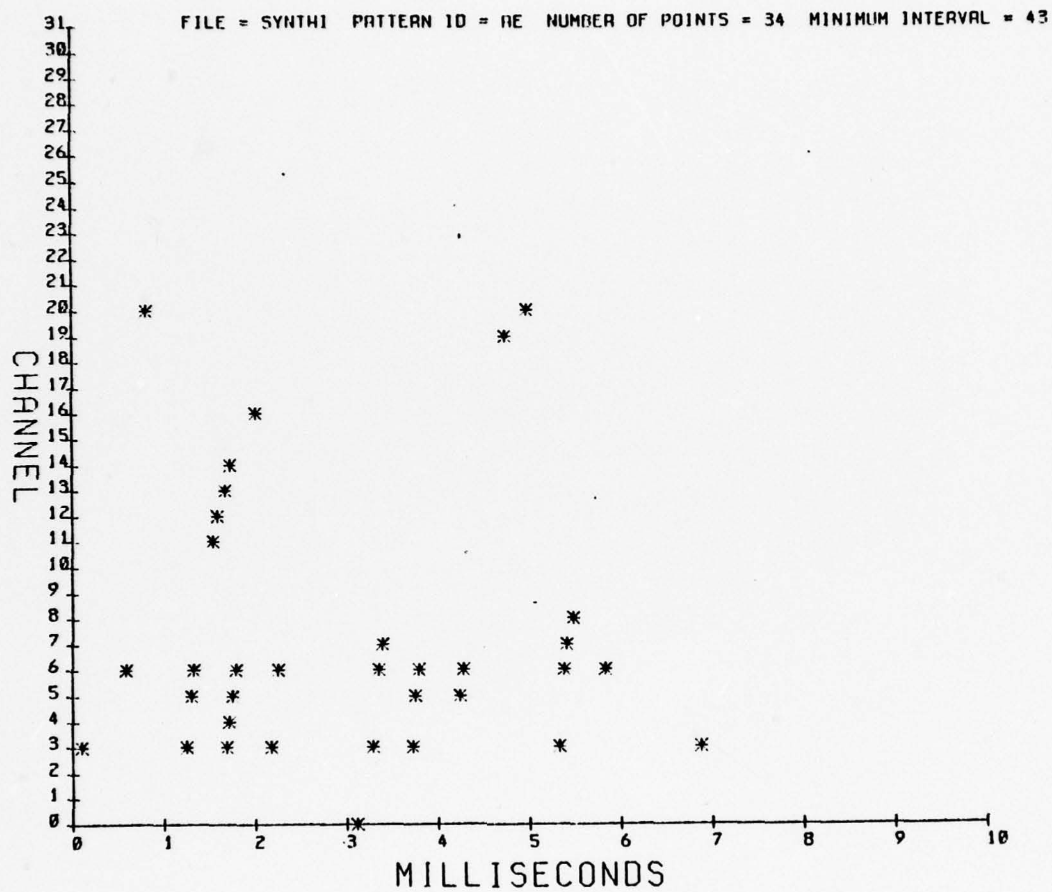

Figure 10. An Example of Reference Pattern.

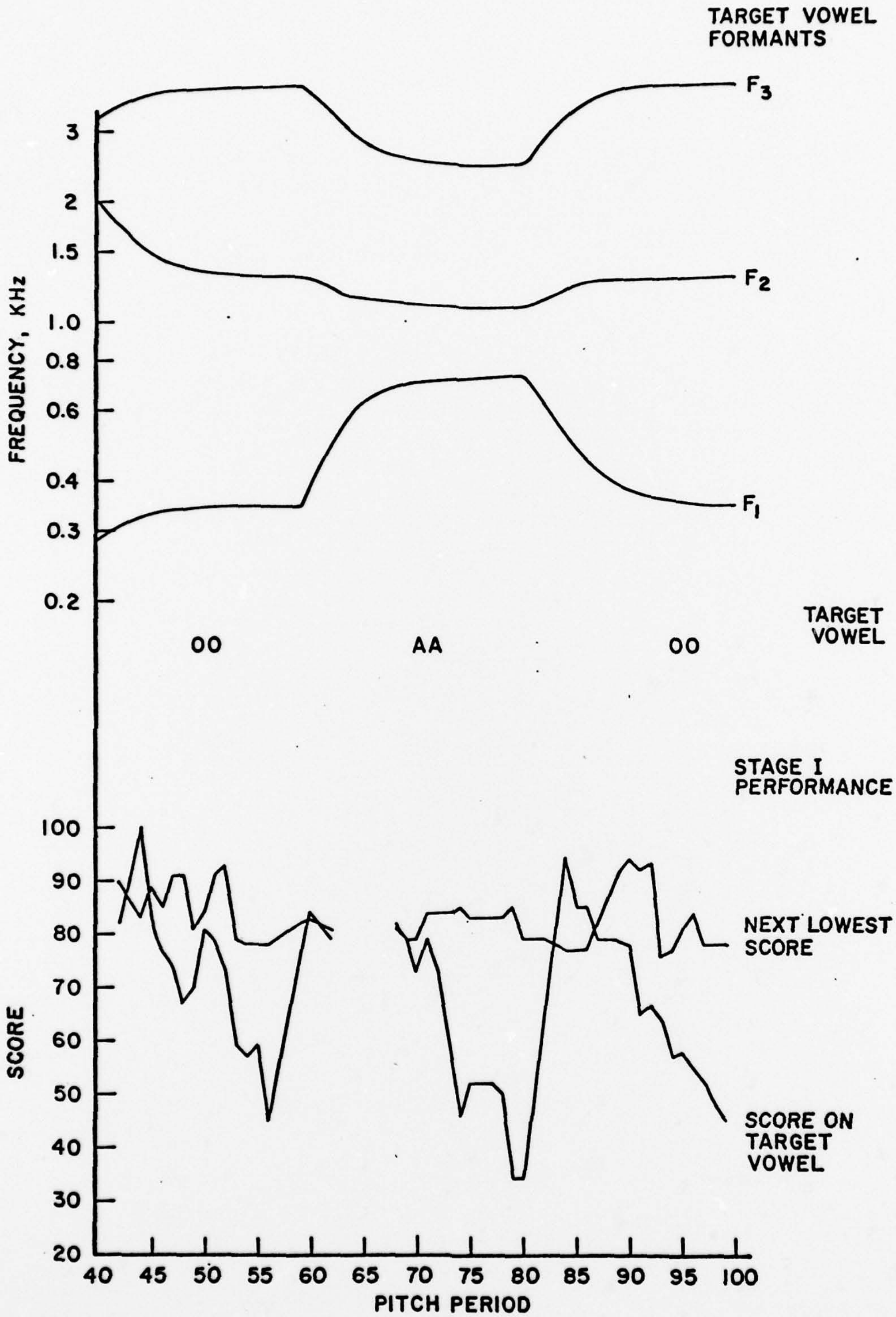

Figure 11. An Example of CxC Stage 1 Performance with QQ as the Stimulus.

vowel. The results clearly suggest that (1) the decision function for Stage I should be based on the lowest score in a pitch period, (2) a pitch period by pitch period decision of phoneme can be made, and (3) positive identification of vowels by Stage II can be based on counting the number of successive pitch periods associated with the same vowel and comparing this number with a threshold.

There was not enough time to quantify the relationship between formant location in the formant frequency space and the pitch period scores, which would have shed some light on what is happening in the time between steady sequences of vowel identifications and provided some guidelines on how close a sound has to be to a vowel before it is identified as that vowel.

5.4.5 Other Incomplete Experiments

In addition to QQ, three other stimuli were constructed and presented to CxC, correlated, and scored, but not processed by the PICKSP program due to lack of time. They were: (1) stimulus containing synthetic voiced stops in a beet-bat environment, (2) a stimulus containing the remaining phonemes associated with the reference patterns, and (3) a real speech stimulus comprised of vowels in an h-d environment.

A detailed evaluation of the results could not be completed during the contract period, but an informal evaluation showed that (1) the system performed as well recognizing the non-vowel phonemes associated with the reference patterns as it did recognizing the vowels, and (2) the system could recognize real speech vowels with no adjustments in the reference patterns.

SECTION 6
DISCUSSION AND RECOMMENDATIONS

The experimental results obtained to date are encouraging. They show that an algorithm exists for the CxC system that enables it to recognize the synthetic phonemes associated with its reference patterns. Of course the evaluation and design are far from complete, and the outcome is far from certain, since there is no assurance that unique trajectories exist in the reference pattern space for the other phonemes.

Probably the most disturbing aspect of CxC to the engineer is the lack of formal understanding of what is happening in the feature extractor portion of the system. Work needs to be done to clarify the principles of operation: What are the features being extracted by the system?

APPENDIX A

DEFINITION OF SYMBOLS

Teletype Code	IPA Symbol	Typical Word
Vowels		
IY	i	beet
II	I	bit
EE	ɛ	bet
AE	æ	bat
AA	ɑ	box
UH	ʌ	but
UU	U	book
OO	u	boot
OW	ɔ	bought
ER	r	bird
Semivowels		
WW	w	word
LL	l	love
RR	r	run
YY	y	yes
Voiced Stops		
BB	b	bat
DD	d	dog
GG	g	got
Voiceless Stops		
PP	p	pot
TT	t	tot
KK	k	cot
Nasals		
MM	m	mat
NN	n	nap
NG	ŋ	sing
Voiced Fricatives		
VV	v	very
TE	ð	the
ZZ	z	zero
ZH	ʒ	azure
Voiceless Fricatives		
FF	f	fine
TH	θ	thick
SS	s	say
SH	ʃ	shoot
Aspirant		
HH	h	help
Affricates		
CH	tʃ	church
JJ	dʒ	judge
Diphthongs		
EI	eI	weigh
AI	aI	tie
OI	ɔI	toy
OU	oU	toe
AU	aU	out

APPENDIX B

AUTOTA SYNTHESIS-BY-RULE

PHONEME CHARACTERISTICS

Phoneme	Formant Frequencies F_1	F_2	F_3	Magnitudes A_V	A_N	Δ1	Δ2	Δ3	Duration
IY	270	2290	3010	100	0	40	40	110	50
II	390	1990	2550	88	0	50	50	90	20
EE	530	1840	2480	60	0	50	55	90	20
AE	660	1720	2410	45	0	40	40	75	50
UH	580	1190	2390	40	0	50	50	50	20
AA	730	1090	2442	38	0	25	40	80	50
OW	570	840	2410	28	0	40	40	80	50
UU	440	1020	2240	43	0	50	50	65	30
OO	350	1300	3900	85	0	40	45	55	50
ER	450	1275	1700	47	0	30	20	30	50
BB	150	600	3000	20	0	50	75	120	20
PP	150	800	1750	0	0	50	40	80	20
MM	280	900	2200	120	0	17	17	40	30
DD	440	2200	3000	20	0	50	50	160	20
TT	440	2200	3000	0	0	50	30	100	10
NN	280	1300	2000	120	0	17	17	100	50
GG	220	1300	1450	20	0	50	50	100	20
KK	220	1300	3300	0	0	50	30	70	20
NG	280	1700	2600	120	0	17	17	100	50
FF	175	900	2400	0	50	20	34	80	20
VV	175	1100	2400	65	35	10	15	40	20
TH	200	1400	2200	0	99	20	28	68	00
TE	200	1600	2200	50	90	10	15	100	00
SS	200	1300	2500	0	40	20	28	50	50
ZZ	200	1300	2500	50	90	20	30	50	20
SH	175	1800	2050	0	99	10	34	100	50
ZH	175	1800	2000	50	40	10	40	100	20
WW	300	610	2200	45	0	25	40	150	00
LL	380	1000	2575	75	0	25	80	150	30
RR	420	1300	1600	50	0				30
YY	300	2200	3065	58	0	25	110	200	00
RO	295	845	1315	80	0	30	80	100	00

REFERENCES

Beek, Bruno, Edward P. Neuberg, and David C. Hodge (1977), "An Assessment of the Technology of Automatic Speech Recognition for Military Applications", IEEE Trans on Acoustics, Speech, and Signal Processing, ASSP-25(4):310.

Hartrum, Thomas (1978), "The CxC Hardware Correlator", Aerospace Medical Research Laboratory, in preparation.

Klatt, Dennis H. (1977), "Review of the ARPA Speech Understanding Project", J. Acoust. Soc. Am., 62(6):1345.

Leet, Duane G. and Christopher W. Walsh (1978), "The CxC System: Standard Operating Procedures for the Analysis of Synthetic and Real Speech, Revision II", UDR-TR-77-58, University of Dayton Research Institute, Dayton, Ohio.

Martin, T.B. and E.F. Grunza (1974), "Voice Control Demonstration System", Technical Report AFAL-TR-74-174, Air Force Avionics Laboratory, Wright-Patterson Air Force Base, Ohio.

Mundie, J.R. (1969), "Neural Calculus", in Biocybernetics of the Central Nervous System, Lorne D. Proctor, Ed., Little, Brown, and Co., Boston.

Mundie, J.R., T. J.C. Rock, and A.J. Goldstein (1974), "Signal Processing Principles Revealed by an Auditory System Model", in Proceedings of the 5th Congress of the Deutsche Gesellschaft für Kybernetik, held at Nuremberg, March 28-30, pp. 292-307.

Mundie, J.R., T. Hartrum, J.C. Rock, and D. Osborne (1978), "The CxC System: General Description", AMRL Report in Processing, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio.

Reddy, D. Raj (1976), "Speech Recognition by Machine: A Review", Proc. of the IEEE, 64(4):501.

Rock, J.C. (1973), "Recursive Plane Analysis: Its Application to the Study of Phase-Locking in Non-Uniform Signal Dependent Sampling Techniques", AMRL-TR-72-34, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio.

Stewart, J.L. (1972), "A Theory and Physical Model for Cochlear Mechanics", Acta Oto-Laryngologica, Supplement 294.

Warmuth, Donald B. (1975), "Speech Synthesis by a Programmable Digital Filter", Air Force Institute of Technology Thesis GE/EE/75-41, Wright-Patterson Air Force Base, Ohio.

Warmuth, Donald B. (1978), "Automatic Recognition of Synthetic Speech Using an Electronic Model of the Middle and Inner Ear", Air Force Institute of Technology Dissertation AFIT/DS/EE/78-3, Wright-Patterson Air Force Base, Ohio.

White, Patrick (1977), "CxC Computer Pattern Classification System Software User's Manual, Revision 1", Systems Research Laboratories, Inc., Dayton, Ohio.

Ziskin, M. and J.R. Mundie (1971), "Encoding Function of Syncoders", AMRL-TR-70-119, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio.